MAP433 Statistique
PC2: Modélisation statistique

5 septembre 2014

1 Stabilisation de la variance

On dispose d'un échantillon $X_1, \ldots, X_n$ i.i.d. de loi de Bernoulli de paramètre $0 < \theta < 1$.

1. On note $\bar{X}_n$ la moyenne empirique des X_i. Que disent la loi des grands nombres et le TCL ?
2. Cherchez une fonction g telle que $\sqrt{n}(g(\bar{X}_n) - g(\theta))$ converge en loi vers Z de loi $\mathcal{N}(0, 1)$.
3. On note z_α le quantile d'ordre $1 - \alpha/2$ de la loi normale standard. En déduire un intervalle $\hat{I}_{n,\alpha}$ fonction de $z_\alpha, n, \bar{X}_n$ tel que $\lim_{n\to\infty} \mathbb{P}(\theta \in \hat{I}_{n,\alpha}) = 1 - \alpha$.

2 Modèle exponentiel

Une grande partie des modèles utilisés en pratique sont des modèles exponentiels (modèle gaussien, log-normal, exponentiel, gamma, Bernouilli, Poisson, etc). Nous allons étudier quelques propriétés de ces modèles. On appelle modèle exponentiel une famille de lois $\{\mathbb{P}_\theta,\ \theta \in \Theta\}$ ayant une densité par rapport à une mesure μ σ-finie sur $\mathbb{R}$ ou $\mathbb{N}$ de la forme

$$p_\theta(x) = c(\theta) \exp\left(m(\theta) f(x) + h(x)\right).$$

On supposera que Θ est un intervalle ouvert de $\mathbb{R}$, $m(\theta) = \theta$ et $c(\cdot) \in C^2$, $c(\theta) > 0$ pour tout $\theta \in \Theta$. On notera X une variable aléatoire de loi $\mathbb{P}_\theta$ et on admettra que

$$\frac{d^i}{d\theta^i} \int \exp(\theta f(x) + h(x))\, \mu(dx) = \int f(x)^i \exp(\theta f(x) + h(x))\, \mu(dx) < +\infty, \quad \text{pour } i = 1, 2.$$

1. Montrez que $\varphi(\theta) := \mathbb{E}_\theta\left(f(X)\right) = -\frac{d}{d\theta} \log(c(\theta))$.
2. Montrez que $\mathrm{Var}_\theta(f(X)) = \varphi'(\theta) = -\frac{d^2}{d\theta^2} \log(c\,(\theta))$.
3. On dispose d'un n-échantillon $X_1, \ldots, X_n$ de loi $\mathbb{P}_\theta$. On note $\hat{\theta}_n$ l'estimateur obtenu en résolvant $\varphi(\hat{\theta}_n) = \frac{1}{n} \sum_{i=1}^n f(X_i)$. En supposant $\mathrm{Var}_\theta(f(X)) > 0$, montrez que

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{\text{loi}} \mathcal{N}\left(0, \frac{1}{\mathrm{Var}_\theta\left(f(X)\right)}\right).$$

3 Modèle probit

Nous disposons d'une information relative au comportement de remboursement ou de non-remboursement d'emprunteurs :

$$Y_i = \begin{cases} 1 & \text{si l'emprunteur } i \text{ rembourse,} \\ 0 & \text{si l'emprunteur } i \text{ est défaillant.} \end{cases}$$

Afin de modéliser ce phénomène, on suppose l'existence d'une variable aléatoire Y_i^* normale, d'espérance m et de variance σ^2, que l'on appellera ≪ capacité de remboursement de l'individu i ≫, telle que :

$$Y_i = \begin{cases} 1 & \text{si } Y_i^* > 0, \\ 0 & \text{si } Y_i^* \leq 0. \end{cases}$$

On note Φ la fonction de répartition de la loi normale $\mathcal{N}(0,1)$.

1. Exprimer la loi de Y_i en fonction de Φ.
2. Les paramètres m et σ^2 sont-ils identifiables ?

4 Modèle d'autorégression

On considère l'observation $Z = (X_1, \ldots, X_n)$, où les X_i sont issus du processus d'autorégression :

$$X_i = \theta X_{i-1} + \xi_i, \quad i = 1, \ldots, n, \qquad X_0 = 0,$$

avec les ξ_i i.i.d. de loi normale $\mathcal{N}(0, \sigma^2)$ et $\theta \in \mathbb{R}$. Écrire le modèle statistique engendré par l'observation Z.

5 Durée de vie

Un système fonctionne en utilisant deux machines de types différents. Les durées de vie X_1 et X_2 des deux machines suivent des lois exponentielles de paramètres λ_1 et λ_2. Les variables aléatoires X_1 et X_2 sont supposées indépendantes.

1. Montrer que une variable aléatoire X suit la loi exponentielle $\mathcal{E}(\lambda)$ si et seulement si
$$\forall x > 0 : \ \mathrm{P}(X > x) = \exp(-\lambda x).$$
2. Calculer la probabilité pour que le système ne tombe pas en panne avant la date t. En déduire la loi de la durée de vie Z du système. Calculer la probabilité pour que la panne du système soit due à une défaillance de la machine 1.
3. Soit $I = 1$ si la panne du système est due à une défaillance de la machine 1, $I = 0$ sinon. Calculer $\mathrm{P}(Z > t; I = \delta)$, pour tout $t \geq 0$ et $\delta \in \{0, 1\}$. En déduire que Z et I sont indépendantes.
4. On dispose de n systèmes identiques et fonctionnant indépendamment les uns des autres dont on observe les durées de vie $Z_1, \ldots, Z_n$.
(a) Écrire le modèle statistique correspondant. Les paramètres λ_1 et λ_2 sont-ils identifiables ?
(b) Supposons maintenant que l'on observe à la fois les durées de vie des systèmes $Z_1, \ldots, Z_n$ et les causes de la défaillance correspondantes $I_1, \ldots, I_n$, $I_i \in \{0, 1\}$. Écrire le modèle statistique dans ce cas. Les paramètres λ_1 et λ_2 sont-ils identifiables ?